Code Generation: Failure of Judgment Analysis

Table 1: **(Java)** False Negatives: analysis of the causes of the misjudgment of the LLMs. Every cell repots the number of occurrences of each category on a sample of 15 cases of misjudgment. Note that more than one category can be assigned to each assignment.

	DSC 1.3B	**DSC 6.7B**	**DSC 33B**	**CL 7B**	**CL 13B**	**CL 34B**	**GPT 3.5**	**GPT 4**
Test result unreliable	1	0	3	0	0	1	3	6
Ambiguous docstring	0	0	1	1	0	0	2	0
Artificial hallucination	7	0	6	11	9	7	5	0
Uncought wrong behavioral	0	0	0	0	0	0	0	0
Misunderstanding of code statements	7	0	3	4	6	7	3	1
Focus on non-functional requirements	0	0	3	0	0	1	0	0
Limited coding context	1	0	0	0	0	2	7	10
Misintepreted implementation requirements	0	0	2	1	0	1	0	1
Shallow description	0	0	0	0	0	0	0	0

Table 2: **(Java)** False Positives: analysis of the causes of the misjudgment of the LLMs. Every cell repots the number of occurrences of each category on a sample of 15 cases of misjudgment. Note that more than one category can be assigned to each assignment.

	DSC 1.3B	**DSC 6.7B**	**DSC 33B**	**CL 7B**	**CL 13B**	**CL 34B**	**GPT 3.5**	**GPT 4**
Test result unreliable	1	0	0	0	0	0	0	1
Ambiguous docstring	4	4	3	1	4	3	4	2
Artificial hallucination	0	0	0	0	0	0	0	0
Uncought wrong behavioral	6	3	10	12	10	7	7	9
Misunderstanding of code statements	0	0	0	0	0	0	0	0
Focus on non-functional requirements	0	0	0	0	0	0	0	0
Limited coding context	9	11	6	1	4	8	7	8
Misintepreted implementation requirements	0	1	0	0	0	0	0	0
Shallow description	0	0	0	0	0	0	0	0

Table 3: **(Python)** False Negatives: analysis of the causes of the misjudgment of the LLMs. Every cell repots the number of occurrences of each category on a sample of 15 cases of misjudgment. Note that more than one category can be assigned to each assignment.

	DSC 1.3B	**DSC 6.7B**	**DSC 33B**	**CL 7B**	**CL 13B**	**CL 34B**	**GPT 3.5**	**GPT 4**
Test result unreliable	3	2	1	1	4	3	3	2
Ambiguous docstring	0	1	2	0	1	2	1	5
Artificial hallucination	1	8	6	4	6	3	4	1
Uncought wrong behavioral	0	0	0	0	0	0	0	0
Misunderstanding of code statements	3	2	0	1	3	2	1	3
Focus on non-functional requirements	0	1	1	0	0	0	0	1
Limited coding context	1	1	2	0	0	1	5	6
Misintepreted implementation requirements	4	0	1	1	0	1	2	1
Shallow description	1	0	1	8	3	3	0	0

Table 4: **(Python)** False Positives: analysis of the causes of the misjudgment of the LLMs. Every cell repots the number of occurrences of each category on a sample of 15 cases of misjudgment. Note that more than one category can be assigned to each assignment.

	DSC 1.3B	**DSC 6.7B**	**DSC 33B**	**CL 7B**	**CL 13B**	**CL 34B**	**GPT 3.5**	**GPT 4**
Test result unreliable	0	0	0	0	0	0	0	0
Ambiguous docstring	7	7	5	7	4	2	6	6
Artificial hallucination	0	0	0	0	0	1	0	0
Uncought wrong behavioral	2	6	4	4	4	8	5	2
Misunderstanding of code statements	0	0	0	0	0	0	0	0
Focus on non-functional requirements	0	0	0	0	0	0	0	0
Limited coding context	5	2	9	3	1	1	4	7
Misintepreted implementation requirements	0	1	0	0	0	0	0	0
Shallow description	1	0	0	2	6	4	0	0